Stat 462/862 Assignment 2

(Due on Oct 22, 2018, hard copy, in the class)

1. Answer the following questions.

   (a) On average, what fraction of people with an odds of 0.4 of defaulting on their credit card payment will in fact default?

   (b) Suppose that an individual has a 18% chance of defaulting on her credit card payment. What is the odds that she will default?

2. Consider the data set *Auto* in the R package *ISLR*. We wish to develop a model to predict whether or not a given car gets high or low gas mileage.

   (a) Create a binary variable, $mpg01$, that contains a 1 if mpg contains a value above this median, and a 0 contains a value below its median.

   (b) Explore the data graphically in order to investigate the association between $mpg01$ and the other features. Which of the other features seem most likely to be useful in predicting $mpg01$? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

   (c) Split the data into a training set and a test set.

   (d) Perform LDA on the training data to predict $mpg01$ using the variables that seemed most associated with $mpg01$ in (b). Report all the parameter estimates. What is the test error of the model obtained?

   (e) Perform QDA on the training data to predict $mpg01$ using the variables that seemed most associated with $mpg01$ in (b). Report all the parameter estimates. What is the test error of the model obtained?

   (f) Perform logistic regression on the training data to predict $mpg01$ using the variables that seemed most associated with $mpg01$ in (b). Report all the parameter estimates. What is the test error of the model obtained?

   (g) Using logistic regression, LDA, and QDA to estimate the probability that a dodge challenger se car with the following setting (cylinders = 6, displacement = 400, horsepower = 110, weight = 3000, acceleration = 15, year = 75, origin = 1) gets high gas mileage.

(h) (Graduate students only) Perform KNN on the training data, with several values of K, in order to predict $mpg01$. Use only the variables that seemed most associated with $mpg01$ in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

3. Consider a dataset (X, Y) in which Y is output and X represents inputs. Let n be the number of observations and p be the number of inputs in the dataset. Consider a special case $n = p = 1$. The ridge regression aims to minimize

$$\sum_{i=1}^{n}(Y_i - \beta_0 - (\sum_{j=1}^{p} X_{ij}\beta_j))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1}$$

while the lasso minimizes

$$\sum_{i=1}^{n}(Y_i - \beta_0 - (\sum_{j=1}^{p} X_{ij}\beta_j))^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

Suppose in both ridge regression and lasso, the intercept is omitted from model.

(a) Choose a few random values of Y and λ, plot (1) and (2) as a function of β, and find their minima on the graphs. Verify that these minima are attached at

$$\hat{\beta}_{ridge} = \frac{Y}{1+\lambda}$$

and

$$\hat{\beta}_{lasso} = \begin{cases} Y - \frac{\lambda}{2}, & \text{if } Y > \frac{\lambda}{2} \\ Y + \frac{\lambda}{2}, & \text{if } Y < -\frac{\lambda}{2} \\ 0, & \text{otherwise.} \end{cases}$$

(b) Choose a few random values of Y, and for each value of Y, plot $\hat{\beta}_{ridge}$ and $\hat{\beta}_{lasso}$ on the same axes, as functions of λ. Describe the observations from the plots.

4. Generate the data $\{X_{i1}, X_{i2}, X_{i3}, Y_i\}_{i=1}^{200}$ using the model $Y_i = 10 + 1.5X_{i1} - 0.3X_{i2} + 10.7X_{i3} + \epsilon_i$, where $X_{i1} \sim Unif(0, 4), X_{i2} \sim Unif(3, 8), X_{i3} \sim Unif(-1, 5), \epsilon_i \sim N(0, 3)$. Now add five more predictor variables $Z_1 = 1.5X_1X_2, Z_2 = -3.6X_1X_3, Z_3 = X_2X_3, Z_4 \sim N(20, 40), Z_5 \sim N(10, 1)$.

(a) Show the Lasso solution path?

(b) What is the tuning parameter that minimizes the cross-validation error? What is the corresponding minimum cross-validation error?

(c) What are the significant variables chosen by the Lasso? Interpret the result.

(d) What is the fitted model?

5. Consider the data set *College* in the R package $ISLR$. The response variable is the number of applications received and the other variables are the predictors.

   (a) Split the data set into a training set and a test set.

   (b) Fit a linear model using least squares on the training set, and report the test error obtained.

   (c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

   (d) Fit a lasso model in the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

   (e) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these three approaches?

6. (Graduate students only) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau I)$, and Gaussian sampling model $y \sim N(X\beta, \sigma^2 I)$. Find the relationship between the regularization parameter λ in the ridge formular, and the variance τ and σ^2.